REMARKS

In the Office Action dated September 20, 2004, the Examiner rejected claims 1-8, 10-12, and 14-17 under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent No. 5,303,327 ("Sturner") in view of U.S. Patent No. 6,253,181 ("Junqua") and in further view of U.S. Patent No. 5,059,127 ("Lewis").

On December 7, 2004 an in-person Examiner Interview was conducted. Participants of the interview included Examiner Armstrong and Applicant's representatives Matthew Sampson and Lisa Schoedel. No exhibits were shown nor demonstrations conducted. The participants discussed claim 1 as well as the Junqua and Lewis references. Further, the participants discussed Applicant's belief that the addition of the Lewis reference to the Section 103(a) rejection does not overcome the deficiencies previously identified in the combination of Sturner and Junqua. As a result of the interview, no agreement with respect to the claims was reached.

Applicant claims a system and method of measuring an ability of a subject. The subject verbally responds to a set of task items, such as the task items depicted in Fig. 2B of Applicant's Specification. A device, such as a speech recognition system, is used to detect the response of the subject and provide an estimate of the speech signal corresponding to the response. However, it is well-known in the art that speech recognition systems have an expected level of accuracy (i.e., speech recognition systems cannot provide an estimate of speech that is 100% accurate). The error behavior of the speech recognition system is item dependent, i.e., different items exhibit different recognizer error patterns. (Specification, page 14, lines 22-23.)

For example, speech recognition systems sample the speech signal as part of the process of estimating speech signals. If the sampling rate is too slow, short words spoken by the subject may be missed and the estimate provided by the recognizer may not accurately reflect the spoken

response. As another example, if a word falls outside the recognizer's predetermined vocabulary, the recognizer is likely to provide an inaccurate estimate of the speech signal. Attached to this response is an article titled "Speech Perception by Humans and Machines" authored by Richard P. Lippmann that compares the ability of speech recognition systems and human ability to accurately recognize speech.

Applicant, recognizing these inherent inaccuracies in speech recognition systems, devised a system and method for providing a subject score that takes into account the expected inaccuracies of a speech recognition system. A scoring computation model is constructed by presenting sets of task items to sample speakers. The sample speakers verbally respond to the task items and their responses are processed by the speech recognition system. By analyzing the speech estimates produced by the recognizer, the inherent inaccuracies of the speech recognition system can be statistically determined and accounted for by the scoring computation model when assessing the ability of a subject. (See, for example, Specification, page 9, line 6 to page 10, line 23 for a description of constructing the scoring computation model.)

Thus, Applicant identifies the inherent inaccuracies in the speech recognition system and creates a scoring computation model of the *expected* task item-dependent operating characteristics of the speech recognition system. The scoring computation model is then used in Applicant's system and method of measuring an ability of the subject. The speech recognition system is not adapted on a user-by-user basis. As a result of Applicant's claimed invention, the subject's ability can be more accurately assessed because the inaccuracies of the recognizer are accounted for in the subject score.

Applicant believes that Sturner, Junqua, and Lewis do not show or suggest a scoring computation model that depends upon "an expected task item-dependent operating characteristic of

the speech recognition system" as claimed. The Office Action states and Applicant agrees that Sturner does not teach that the subject score accounts for the item-dependent operational characteristics of the speech recognition system. (See, Office Action, page 3.) Applicant also believes that Junqua and Lewis do not teach a scoring computation model that depends upon an expected task item-dependent operating characteristic of the speech recognition system.

In contrast to Applicant's claimed invention, Junqua describes adapting a speech recognition system on a user-by-user basis. Speech recognition systems can be classified as being either speaker-dependent technology or speaker-independent technology. Speaker-dependent technology requires a user to adapt the speech recognition system to recognize the particular speech patterns of a particular person, while speaker-independent technology requires no training. While, speaker-dependent systems are more accurate once they have been trained, they are not appropriate for applications in which there is no predetermined user(s) associated with the speech recognition system. For example, applications in which the general public accesses the speech recognition system typically use speaker-independent technology.

Junqua's teachings are directed towards speaker-dependent technology. As described in Junqua's background section:

> [T]he invention relates to a speech recognition apparatus having an adaptation system employing eigenvoice based vectors to rapidly adapt the initial speech model to that of the user. The system further employs a confidence measuring technique whereby the system automatically bases its adaptation upon utterances recognized with high confidence, while ignoring utterances recognized with low confidence. In this way, the system automatically adapts to the user quite rapidly, increasing the recognizer's chance of having a good recognition performance, without adapting to incorrect pronunciations.

(See, Junqua, column 1, lines 10-20.) As described above, use of Junqua's apparatus results in a modification to the initial speech model of the speech recognition apparatus. Unlike Junqua, Applicant makes no modifications to the initial speech model or any other component of the speech

recognition system on a user-by-user basis. Applicant instead identifies the inherent inaccuracies in the speech recognition system and then creates a scoring computation model of the expected task item-dependent operating characteristics of the speech recognition system to be used in Applicant's claimed system and method for measuring an ability of a subject.

Moreover, Junqua is silent with respect to inherent inaccuracies of the speech recognition system. Junqua describes overcoming problems associated with inaccurate responses, but not inaccuracies associated with the recognizer itself. In contrast to Junqua's silence regarding the item-dependent operating characteristics of a speech recognition system, Applicant describes this problem and its affect upon measuring an ability of a subject. The applicant then describes and claims a system and method of measuring an ability of the subject that accounts for the item-dependent characteristics of the recognizer.

Lewis fails to overcome the deficiencies in Sturner and Junqua. Lewis describes a computerized implementation of sequential testing that is based on Item Response Theory. (See, e.g., Lewis, Abstract.) However, Lewis is silent with respect to speech recognition systems. Because Lewis is silent with respect to speech recognition systems in general, it follows that Lewis is more specifically silent with respect with the inherent inaccuracies of speech recognition systems.

Sturner, Junqua, and Lewis are all silent with respect to the item-dependent characteristics (e.g., inaccuracies) of a speech recognition system. At most, Junqua teaches an apparatus that describes the response-dependent adaptation of a speech recognition system. Thus, the combination of Sturner, Junqua, and Lewis would not result in a system in which a scoring computation model is based upon "an expected task item-dependent operating characteristic of the speech recognition system" as claimed. Because Sturner, Junqua, and Lewis do not show or suggest a scoring computation model that depends upon an expected task item-dependent operating

characteristic of the speech recognition system, Applicant submits that claims 1, 7, 8, 14, 16, and 17 are not obvious in light of the combination of Sturner, Junqua, and Lewis.

Claims 2-6 depend on claim 1. Claims 10-12 depend on claim 8. Claim 15 depends on claim 14. Accordingly, Applicant also submits that claims 2-6, 10-12, and 15 are allowable for at least the reasons set forth above.

In light of the above, Applicant respectfully requests withdrawal of the rejections under 35 U.S.C. § 103(a).

CONCLUSION

In light of the above remarks, Applicant submits that the present application is in condition for allowance and respectfully requests notice to this effect. The Examiner is requested to contact Applicant's representative below if any questions arise or she may be of assistance to the Examiner.

Respectfully submitted,

Date: December 20, 2004

By: 

Lisa M. Schoedel
Reg. No. 53,564
McDonnell Boehnen Hulbert & Berghoff LLP
300 South Wacker Drive
Chicago, Illinois 60606-6709
312 935 2362
schoedel@mbhb.com

Presented at the European Speech Communication Association Tutorial and Research Workshop on "The Auditory Basis of Speech Perception," Keele University, UK, July 15-19, 1996

SPEECH PERCEPTION BY HUMANS AND MACHINES*

Richard P. Lippmann
email: rpl@SST.LL.MIT.EDU
Room S4-121, Lincoln Laboratory MIT
244 Wood Street
Lexington, MA 02173-9108
USA

ABSTRACT

This paper reviews past research on human speech perception and recent studies which compare the performance of humans and speech recognizers using six modern speech corpora with vocabularies ranging from 10 to 65,000 words. Error rates of machines are often more than an order of magnitude greater than those of humans for quiet, clearly spoken speech. Machine performance degrades further below that of humans in noise and under other stressing conditions. Human performance remains high with natural variability caused by new talkers, spontaneous speaking styles, noise, and reverberation. Human performance also remains high with unnatural degradations caused by waveform clipping, band-reject filtering, and analog waveform scrambling. Humans can also recognize quiet, clearly spoken nonsense syllables and words without high-level grammatical information. Much further algorithm development is required before even the low-level acoustic-phonetic accuracy of machines equals that of humans on real-world tasks.

1. INTRODUCTION

Dramatic advances have been made in speech recognition technology over the past few years and commercial recognizers are being widely applied in a number of limited application areas. It is likely, however, that recognizers will not become commonplace and enjoy widespread use until performance approaches that of humans. This paper brings together results from many scattered studies which compare human and machine speech recognition to determine how much speech recognizers must improve to match human performance. In addition, results of speech perception experiments are presented which demonstrate the resistance of humans to both natural and unnatural speech variability and degradations. The primary motivations for this paper are to promote a multi-disciplinary dialog between speech recognition and speech perception researchers, and to determine new research directions for both speech recognition and speech perception which can help close the gap between human and machine performance. The remainder of this paper first presents human/machine comparisons using six large talker-independent speech corpora. Human speech perception results obtained across talkers, with filtering, in noise, and with nonlinear degradations are then presented, followed by a summary and discussion.

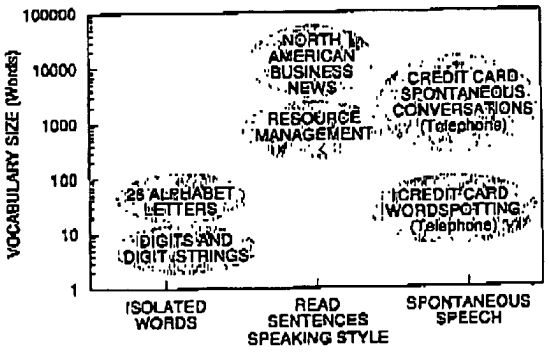

FIGURE 1. Six talker-independent speech recognition corpora used to compare humans and machines.

2. RECENT HUMAN-MACHINE COMPARISONS

Human and machine speech recognition performance is compared in this section using the six speech corpora shown in Figure 1. These corpora span a wide range of difficulty and represent many different potential applications of speech recognition technology. All are designed to test talker-independent recognition of speech from talkers not used for training. Vocabulary sizes range from 10 to 65,000 words and materials are

*This work was sponsored by the Department of Defense Advanced Research Projects Agency. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

produced using a wide range of speaking styles. Materials were recorded by prompting talkers to produce words in isolation, by having talkers read carefully prepared sentences, and by recording extemporaneous telephone conversations on the topic of "credit cards."

All corpora are used for a dictation task where the goal is to identify all the words which were spoken. Word error rates treat substitutions, deletions, and word insertions as errors. The credit-card corpus is also used to evaluate the ability of wordspotters to detect 20 keywords in conversational telephone speech. Wordspotters are useful for computer and telephone interfaces with untrained users because they do not rely on constraining grammars to provide good performance. The performance metric used for wordspotting is the average detection rate, or the percentage of true keyword occurrences that are detected. The "miss" rate is 100 minus the detection rate. The detection rate is averaged over wordspotter systems adjusted to provide false alarm rates ranging from 1 to 10 false alarms per keyword per hour of conversational speech.

Corpus	Type	Talkers	Words	Utterances	Duration	Perplexity
TI Digits	Digits	326	10	25,102	4 hrs	11
Alphabet Letters	Alphabet Letters	150	26	7,800	1 hr	26
Resource Management	Sentences	109	1,000	4,000	4 hrs	60-1,000
North Amer. Business News (NAB)	Sentences	84 - 284	5,000 - 20,000	7,200 - 37,200	12 hrs - 62 hrs	45-160
Credit-Card CSR	Conversations	70	2,000	35 Convs., 1,600 Segs.	2 hrs	100
Credit-Card Wordspotting	Conversations	70	20 Keywords	2,000 Occurrences	2 hrs	-

TABLE 1 Characteristics of six speech corpora.

Characteristics of the six speech corpora are provided in Table 1. All but the Credit-Card corpora contain read speech and this table summarizes information concerning all the materials in a corpus (TI Digits, Alphabet Letters) or concerning the data available for training (all other corpora). The "Words" column refers to the vocabulary size, the "Utterances" column refers to the number of sentences, words, conversation fragments, or keyword occurrences that were recorded, and the "Perplexity" column measures the average number of words that can occur following any other word. In this table, perplexity is a measure of the word-sequence constraints provided by the recognition grammars used for each speech corpus. It is a better indicator of difficulty than vocabulary size.

These corpora include a wide range of speech materials. The TI-digits corpus [16] contains isolated

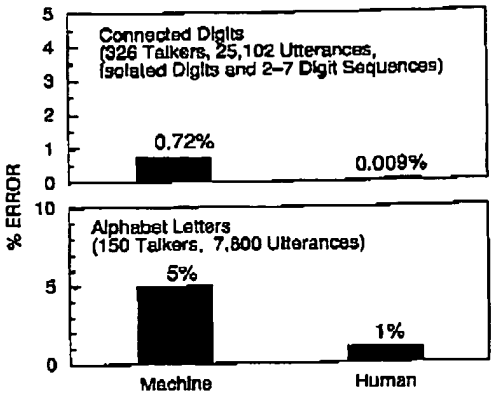

FIGURE 2. Human and machine error rates for connected digits and isolated alphabet letters.

digits and two-to-seven digit sequences. The perplexity of this task is eleven because the digit "*O*" can be pronounced "*zero*," or "*oh*." The alphabet letters corpus [4] contains isolated spoken letters which could be used to spell a person's name. The Resource Management corpus [31] contains highly constrained sentences that can be used to query a naval data base. Two sample sentences are "*Are there fourteen ships at sea*," and "*List all cruisers and their fleet identifications.*" The Wall Street Journal corpus [30] contains sentences selected from articles in the Wall Street Journal. Two sample sentences are "*For the first time in years the Republicans also captured both houses of Congress*," and "*Terms weren't disclosed, but industry sources said the price was about $2.5 million.*"

The Credit Card component of the NIST Switchboard corpus [14] is used for both wordspotting and continuous speech recognition (CSR). It contains speech recorded over normal telephone lines from two talkers located at home or work carrying on a spontaneous conversation concerning credit cards. Two sample excerpts are "*I don't know if I'm really afraid of spending too much*," and "*I uh, I try to get maybe just one or two.*" These phrases contain false starts and are frequently non grammatical due to the spontaneous nature of the conversations. This material also samples a wide range of talking styles and acoustic environments. Wordspotter tests with this corpus search for 20 frequently occurring keywords including "card," "credit," and "charge." Continuous speech recognition tests recognize all spoken words.

Human responses for the digit and alphabet corpora were recorded by typing words from a closed response set. Figure 2 shows human and machine error rates for these two corpora. Results for digit strings are shown in the top of this figure. Human results are average digit string error rates obtained using wide-band listening tests

and highly motivated listeners as described in [16]. The machine digit string recognition error rate is the average digit string error from [3] obtained using a hidden Markov model (HMM) recognizer designed specifically for this task. It has the lowest error rate reported to date on this corpus. The upper part of Figure 2 demonstrates that highly motivated human performance on connected digits is extremely good. The error rate of 0.009% represents an average of 2-3 errors per listener over more than 25,000 tokens. Machine performance at 0.72% is almost two orders of magnitude worse.

Human and machine error rates for spoken letters of the alphabet are presented in the lower part of Figure 2. The 1% human error rate for this corpus [7] is similar to the 1-2% human error rates that have been reported for consonant-vowel-consonant (CVC) nonsense syllables [8][19]. The machine recognition error rate is from a neural network recognizer designed specifically for recognizing isolated letters [4]. It has the best performance reported to date on this task. As can be seen, the 5% machine error rate is five times higher than the 1% human recognition error rate.

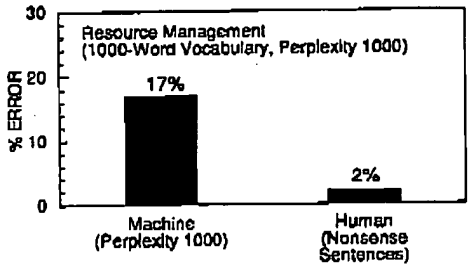

FIGURE 3. Machine error rates for the Resource Management corpus with a null grammar (perplexity of 1000) and human error rates for nonsense sentences.

No formal listening tests were performed for the Resource Management speech corpus. Instead, Resource Management error rates obtained with a "null grammar" can be compared to human results obtained with nonsense sentences. A null grammar assigns equal probability to all words and results in a perplexity of 1000. This grammar used in a speech recognizer is roughly equivalent to nonsense sentences used with human listeners. The null grammar recognition condition uses a recognizer which makes no use of word sequence information, while nonsense sentences provide human listeners with limited word sequence information. Nonsense sentences are semantically meaningless word sequences created by randomly selecting four or five keywords from a list of 1,000 keywords and placing them in keyword slots within fixed sentence frames. Examples of these sentences with sample keywords shown using italics include "The *cuff golf* told the *hold dive*," and "A

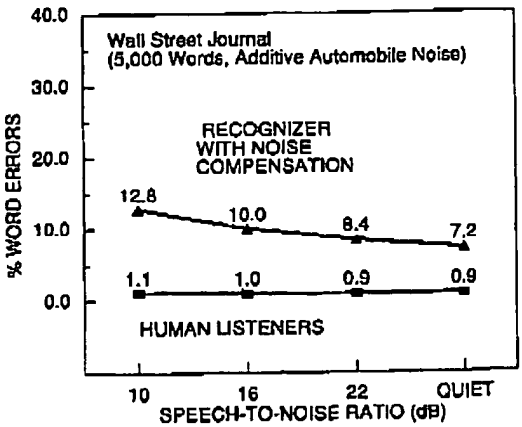

FIGURE 4. Performance of humans and of a HMM recognizer with noise compensation for Wall Street Journal sentences with additive automobile noise.

golden corner varies the *thoughtful keeper*." Words in these sentences are produced at a faster rate than in isolation and with cross-word coarticulation, but with little contextual information.

Figure 3 compares error rates obtained using the best performing HMM recognizer on the Resource Management corpus [11] to human error rates for nonsense sentences [19][21]. Human error rates were obtained when listeners knew words were selected at random, with an unlimited response vocabulary, and by providing roughly 10 seconds to write responses. Machine error rates are for a limited 1000 word vocabulary using an HMM recognizer that was highly tuned to this corpus. The 17% machine word error rate is almost an order of magnitude higher than the 2% human word error rate. These results demonstrate that humans can accurately perceive wideband speech in a quiet environment with no contextual information using only low-level acoustic phonetic information. The representative high-performance HMM recognizer described in [11], however, has poor low-level acoustic modeling and relies heavily on contextual information and restrictive grammars to achieve good performance. For example, its error rate drops to 3.6% when perplexity is reduced from 1000 to 60 using a highly constraining word-pair grammar [29]. Informal tests with one listener [36] suggest that human error rates are reduced to 0.1% with context under the same low-perplexity condition.

Human error rates are reported in [6] for Wall Street Journal sentences from one condition (Spoke 10) of the 1994 ARPA continuous speech recognition evaluation which evaluated recognition performance in noise [28]. The Wall Street Journal corpus used in this evaluation is now considered part of the North American Business

News (NAB) corpus shown in Figure 1. Error rates of recognizers were evaluated in quiet and at three speech-to-noise ratios (SNRs) using a 5,000 word vocabulary. Noise was recorded in an automobile and added to sentences which were recorded with a high-quality close-talking microphone in quiet. Segments containing only noise were provided for training HMM recognizers and developing new algorithms to adapt recognizers to the noise environments. Figure 4 shows word error rates for an adaptation algorithm described in [9] which provides good performance and requires only a few seconds of noise-only adaptation data. Figure 4 also shows average human performance after obvious typing errors, spelling errors and out-of-vocabulary words were corrected. Humans recognize sentences with a word error rate of roughly 1% even with noise at the lowest SNR of 10 dB. The error rates of recognizers with extensive adaptation, however, increased substantially with noise that had little effect on human listeners. Machine error rates range from 7.2 to 12.8% and are roughly ten times worse than human performance. Error rates without adaptation can not be plotted on the scale used in Figure 4 because they increase to 42.2% and 77.4% at the lowest two SNRs. Similar results were obtained in a separate study [15] which compared human and machine performance using Wall Street Journal sentences. In these and other experiments, the error rates of machine recognizers increase substantially at noise levels which do not even affect human listeners.

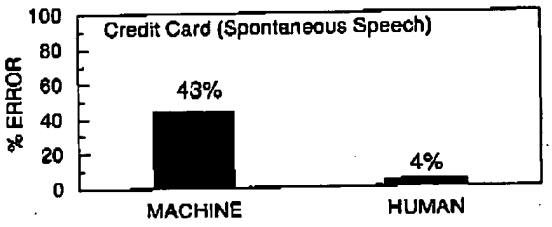

FIGURE 5. Word error rates for humans and a HMM recognizer on phrases extracted from spontaneous telephone conversations in the credit card speech corpus.

The accuracy of transcriptions for Credit Card telephone conversations provides a measure of human performance for these materials. These transcriptions were created by court reporters and temporary employees, and include non-speech sounds such as coughs, laughs, and breath noise [14]. The accuracy of over 14,000 transcribed words was carefully validated by linguists and speech scientists using repeated listening to both sides of each conversation, examinations of the waveform, and group-vote consensus decisions for difficult utterances which were faint, spoken rapidly, or partly masked by simultaneous speech from the opposite talker. The average transcription error rate, counting insertions, deletions, and substitutions, was 4% [22].

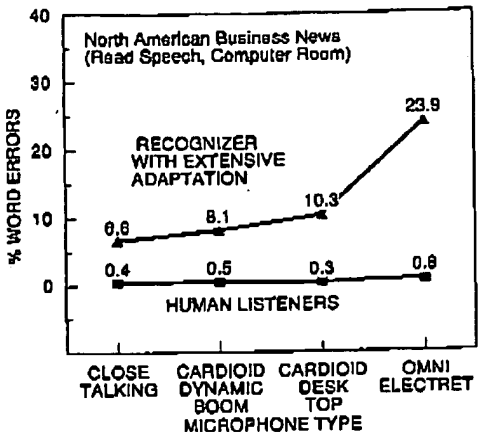

FIGURE 6. Human and machine error rates with multiple microphones for the NAB corpus.

Speech recognition performance on this corpus is summarized in Table 8 of [20] for an HMM recognizer that provides state-of-the-art performance on this corpus. Figure 5 compares word error rates of human transcribers and of this HMM recognizer.

The error rate of the HMM recognizer on the spontaneous speech Credit Card corpus is roughly 43%, which is extremely high compared to the much lower rates obtained with the Resource Management and Wall Street Journal corpora using read speech. This large increase in error rate is caused by many factors including the talking style and weak grammar used for conversations. Results obtained with other read speech corpora suggest that the limited bandwidth provided by the telephone system accounts for only a small component of the increase [28]. This result again demonstrates the lack of robustness of current speech recognition technology. Error rates increase dramatically for spontaneous speech under conditions where human performance remains high. The 43% machine error rate on this corpus is almost an order of magnitude greater than the 4% human error rate.

Large vocabulary HMM speech recognizers were recently compared to humans using NAB corpus sentences and speech recorded using four different microphones in a slightly reverberant office environment with background acoustic noise which resulted in SNRs of roughly 20 dB. Speech materials were recorded with the same close-talking microphone used for the Resource Management and Wall Street Journal evaluations, with two high-quality cardioid studio microphones, and with a low-cost omni-directional electret microphone. Figure 6 shows machine error rates for a 65,000 word HMM recognizer that obtained the best results for this task [33]. Also shown are human results obtained using a majority

vote from the responses of three listeners to eliminate errors due to inattention [5]. Human error rates vary little and are below 0.8% for all conditions. Machine error rates increase dramatically from 6.6% with the close talking microphone to roughly 24% with the electret microphone. This increase occurs despite extensive adaptation algorithms which were used to compensate for microphone variability. Machine error rates under all conditions are more than an order of magnitude greater than human error rates and machine performance degrades dramatically under conditions where human performance degrades only slightly.

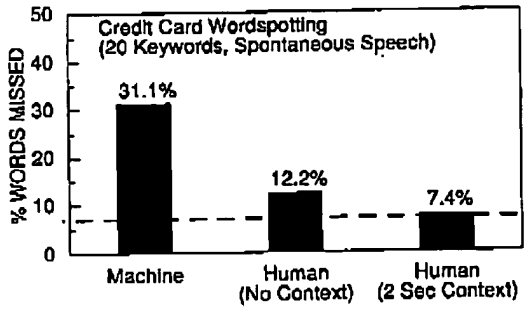

FIGURE 7. Average human and machine miss rate for 20 keywords in the credit card speech corpus.

Figure 7 compares human and machine wordspotting performance on the Credit Card corpus using data from [2]. True occurrences of each of 20 keywords and an equal number of false alarms were presented at random to listeners. False alarms were generated by a high-performance, whole-word, hybrid HMM/ neural-network wordspotter which provides the best null-grammar performance reported on this corpus. Listeners determined whether a presented utterance was a designated keyword. Two listening conditions were used to evaluate the importance of context. In a "no context" condition, listeners were presented false alarms and keyword occurrences with 100 msecs of extra speech both before and after each segment. This small amount of extra context was necessary to keep from chopping off the beginnings and end of keywords. In the "2 second context" condition, all sounds beginning two seconds before and ending two seconds after each utterance were presented. This context typically included most of the phrase containing each utterance. Resulting human judgements were used to compute new average detection and miss rates for each word by eliminating all utterances judged to be non-keywords. The highest human detection rate that can be achieved in this experiment is not 100% because humans did not listen to all wordspotter false alarms and thus could not eliminate all false alarms. The highest detection accuracy is 92.8% and the lowest miss rate is thus 7.2% as shown by the dotted line in Figure 7.

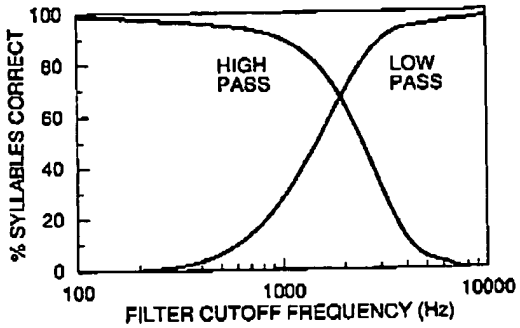

FIGURE 8. Human nonsense syllable accuracy with sharp high-pass and low-pass filtering.

Figure 7 shows that human judgements reduce the average miss rate for keywords from 31.1% for the wordspotter to 12.2% for humans with no contextual information and 7.4% for humans with two seconds of contextual information. The low human miss rate without context demonstrates that humans can make fine phonetic distinctions between similar sounding words such as "car" and "card" that are difficult for current recognizers. Human performance with two seconds of context indicates almost perfect discrimination between true keyword occurrences and false alarms with limited context.

3. HUMAN SPEECH PERCEPTION

Although machine recognition accuracy degrades rapidly with noise and channel variability, humans are able to recognize speech with a surprising amount of noise, filtering, and distortion and we can tolerate a large amount of variability across talkers and talking style. Figure 8 shows CVC nonsense syllable accuracy for well-trained crews of human listeners used in studies performed at Bell Laboratories in the late 1920's [8]. Speech syllable intelligibility for this difficult task remains above 90% for either high-pass filtering above 1 kHz or low-pass filtering below 3 kHz. Speech intelligibility drops only slowly as additional filtering is applied and although filtered speech sounds somewhat different from wideband speech, little training is required to recognize filtered speech accurately.

Figure 9 shows error rates for four other types of linear filtering. The left-hand bar shows that word error rates remain less than 3% when the speech spectrum is tilted up or down by a differentiator with a frequency response that rises 6 dB/octave or an integrator with a frequency response that falls by 6 dB/octave [17]. The middle bar shows that the error rate for words in sentences remains below 10% even when speech is presented through a highly erratic frequency response formed from three separate 500 Hz wide sharp bandpass

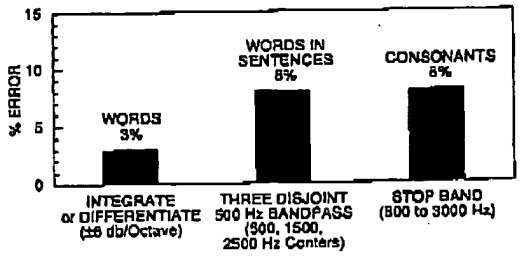

FIGURE 9. Intelligibility of speech materials for humans with linear filtering channel variability.

filters with centers at 500, 1500, and 2500 Hz [13]. The right bar shows that consonant error rates remain below 10% even when speech energy from 800 Hz to 3000 Hz is eliminated using sharp high-pass and low-pass filters [18]. In this condition, listeners must integrate speech cues from below 800 Hz and above 3000 Hz which are normally redundant with acoustic cues conveyed by mid-frequency speech energy. These results demonstrate that human listeners are able to perceive speech accurately with widely varying linear frequency responses and with little or no training under unnatural novel listening conditions.

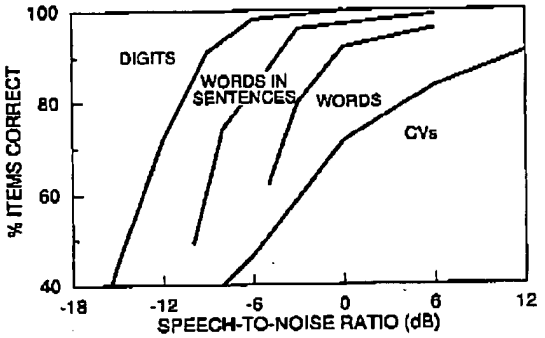

FIGURE 10. Intelligibility of speech materials presented to human listeners with additive noise.

Speech intelligibility decreases gradually when masking noise is added at progressively higher levels and as the amount of contextual information available to a listener decreases. Figure 10 summarizes the results from three representative experiments that explored the effects of additive noise. These experiments measure error rates for spoken digits in wide-band noise [24], for Harvard sentences and PB-50 words in speech-shaped low-pass filtered noise [34], and for consonant-vowel (CV) nonsense syllables in wide-band noise [23]. As can be seen, performance degrades only slowly as noise is added. Intelligibility of digits, isolated words, and words in sentences is well above 90% correct even at a speech-to-noise ratio (SNR) of 0 dB. Speech intelligibility degrades slowly as the SNR becomes lower

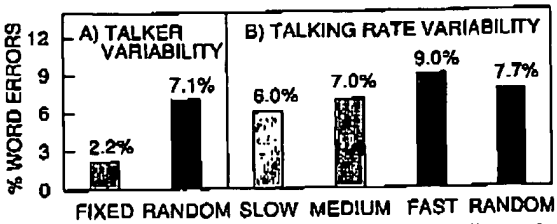

FIGURE 11. Word error rates (A) when the talker and (B) when the talking rate is held fixed across a word list or varied randomly from token-to-token within a list.

until accuracy for all materials is below 50% correct when the SNR is below -15 dB. These results are representative. Actual error rates depend strongly on the noise spectrum, the technique used to measure SNR, and other variables common to this type of testing (e.g. talkers, listener training, system bandwidth, and vocabulary size). In general, however, when noise with a flat or slowly falling spectrum is used, speech intelligibility for words in sentences and for isolated words does not drop substantially until the SNR drops below 0 dB.

Human listeners are tolerant of variability across both talkers and talking rate. Word error rates measured in the quiet with wideband speech are almost always below 10% across talkers and listeners. The effects of talker variability are normally averaged out in many intelligibility tests which sample speech from only a few talkers, and present materials from one talker at a time. A more stressing condition, which evaluates the ability of listeners to rapidly adapt to new talkers, uses speech produced from many talkers and varies the talker who produced a word after each word token is presented. The left side of Figure 11 shows results from this condition in a recent experiment [27] which used quiet wideband speech and either 1 or 15 talkers. Average word error rates for results in quiet when the talker was held fixed throughout each word list were 3.2%. These error rates increased by less than 5 percentage points to 7.8% when the talker was varied randomly for each token in a word list. This and other similar studies demonstrate that error rates do not vary widely across talkers for quiet wideband speech.

Talkers make frequent and extensive changes in speaking rate during normal conversations and human listeners typically maintain high speech intelligibility during these changes. The right side of Figure 11 presents recent results [32] obtained when talkers produce isolated monosyllabic words at a normal rate (533 msecs per word or 113 words/minute) at a fast rate (1.4 times normal or 160 words/minute) and at a slow rate (0.6 times normal or 66 words/minute). Word error scores increase only slightly by 3 percentage points from 6% to 9% when the talking rate increases. Also show in the right of Figure 11

is the word error rate when talking rate is varied randomly across these three rates for every token in a word list. This 7.7% error rate with token-to-token talker variability is only slightly higher than the average 7.0% error rate obtained when the talking rate is constant across a word list. These results demonstrate that error rates for normal listeners are low, and do not vary widely, as talking rate is varied over a wide range.

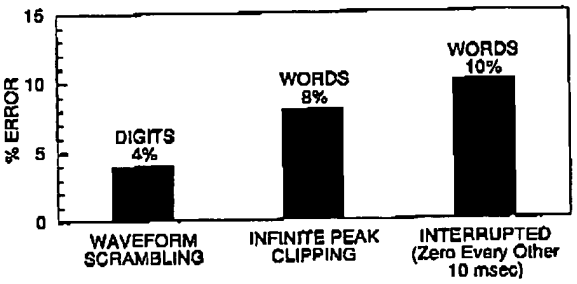

FIGURE 12. Human word error rates with three types of nonlinear distortions.

Other experiments in human speech perception demonstrate that speech intelligibility remains high with many types of nonlinear distortions. Historically, this has been most noted by engineers who have tried to build analog speech scramblers for telephone privacy. It was found that complex systems which scramble the speech waveform and spectrum, often degrade intelligibility little. An evaluation of the intelligibility of digits processed by various waveform and spectral scramblers is reported in [12]. One waveform scrambler chopped the speech waveform into 8 to 32 msec frames and then randomized the presentation order of every 16 to 4 sequential frames. This preserves local spectral information within a frame but distorts temporal sequencing. Digit error rates for this type of waveform scrambling were never greater than 4% as shown in the upper left of Figure 12.

Figure 12 shows that speech intelligibility remains surprisingly high with other types of spectral distortions. The middle bar shows that the word error rate remains below 8% even when speech is differentiated and then processed using infinite peak clipping. This preserves only the zero crossings of the speech waveform and results in a constant amplitude waveform [17]. The word error rate with infinite peak clipping drops to roughly 2% when human listeners are provided training over 15 sessions. The right bar in Figure 12 shows that the error rate of PB-50 words does not rise above 10% even when every other 10 to 40 msec segment of the speech waveform is zeroed out [25]. This condition eliminates 50% of the speech waveform. These results with distortions demonstrate that speech has many redundant cues and that humans can ignore distorted or missing cues, even with highly unnatural distortions, and focus on remaining important cues.

4. SUMMARY AND DISCUSSION

Machine speech recognition word error rates are often more than an order of magnitude higher than those of humans in quiet environments. Machine performance degrades even further below that of humans in noise, with channel variability, and with spontaneous speech. Variability caused by noise and by training and testing with different microphones, which severely degrades machine recognition accuracy, often has little or no affect on humans. Talker and talker-rate variability for quiet speech also has little effect on humans. Humans have excellent low-level acoustic-phonetic perception and can recognize words extracted from spontaneous conversations, words in nonsense sentences, and CVC nonsense syllables without high-level grammar information.

Humans require no prior training to recognize speech with highly unnatural types of distortions including waveform clipping, band-reject filtering, and waveform scrambling. We are able to integrate diverse spectral, temporal, and suprasegmental cues and understand speech with many types of natural and unnatural distortions. Machine recognizers provide best performance with degraded speech only when they are trained using degraded speech materials or when internal parameters are adapted to mimic this type of training. Humans do not require retraining for every new situation.

These results, and other results on human perception of distorted speech suggest that humans are using a process for speech recognition that is fundamentally different from the simple types of template matching that are performed in modern speech recognizers. Humans appear to be able to ignore missing or distorted cues and focus on remaining important cues. They are also much less sensitive to channel variability. This suggests two important research directions. The first is to develop new approaches to extracting features derived from narrow spectral/temporal regions that will be less sensitive to channel variability. The second is to add active analysis in the front-ends of speech recognizers to determine when a feature is present and also when it is a component of a desired speech signal. This supplementary information can be used by classifiers that can compensate for missing features. Some promising preliminary work with narrow band features has recently been presented [26] and researchers are exploring the ability of auditory scene analysis and speech activity detection to identify missing features (e.g. [10]). In addition, this review suggests that further studies should explore machine performance with nonlinear distortions and extreme types of channel

variability, and evaluate the time course of human adaptation to filtering, noise, and other degradations.

5. REFERENCES

1. H. Bourlard, H. Hermansky, and N. Morgan, "Towards Increasing Speech Recognition Error Rates," Speech Communication, 18(3), 1996.
2. E. Chang, and R. Lippmann, "Improving Wordspotting Performance with Artificially Generated Data", ICASSP, 526-529, 1996.
3. W. Chou, C. H. Lee, and B. H. Juang, "Minimum Error Rate Training of Inter-Word Context Dependent Acoustic Model Units in Speech Recognition", ICSLP, Yokohama, Japan, S09:3.1-3.4, 1994.
4. Ronald Cole, Mark Fanty, et al., "Speaker-Independent Recognition of Spoken English Letters," 1990 IEEE INNS International Joint Conference on Neural Networks, Vol. 2, 45-51, San Diego, CA, 1990.
5. N. Deshmukh, A. Ganapathiraju, et al., "Human Speech Recognition Performance on the 1995 CSR Hub-3 Corpus," SLST Workshop, Harriman NY, Morgan Kaufmann, 1996.
6. W. J. Ebel and J. Picone, "Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus," SLST Workshop, Austin TX, 53-59, Morgan Kaufmann, 1995.
7. Mark Fanty, Philipp Schmid, and Ronald Cole, "City Name Recognition Over the Telephone," ICASSP, 549-552, 1993.
8. N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," JASA, 19, 90-119, 1947.
9. R. A. Gopinath, M. Gales, P. S. Gopalakrishnan, S. Balakrishnan-Aiyer, and M. A. Picheny, "Robust Speech Recognition in Noise - Performance of the IBM Continuous Speech Recognizer on the ARPA Noise Spoke Task", SLST Workshop, Austin TX, 127-130, Morgan Kaufmann, 1995.
10. P. D. Green, M. P. Cooke, and M. D. Crawford, "Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise," ICASSP, 401-404, 1995.
11. X. D. Huang, K. F. Lee, H. W. Hon, and M. Y. Hwang, "Improved Acoustic Modeling with the SPHINX Speech Recognition System", ICASSP, 345-348, 1991.
12. Jayant, N.S., et al., "A Comparison of Four Methods for Analog Speech Privacy," IEEE Transactions on Communications, COM-29(1), 18-23, 1981.
13. Karl D. Kryter, "Speech Bandwidth Compression through Spectrum Selection," JASA, 32(5), 547-556, 1960.
14. SWITCHBOARD: A User' s Manual, Catalog Number LDC94S7, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.
15. D. A. van Leeuwen, L. G. van den Berg, and H. J. M. Steeneken, "Human Benchmarks for Speaker Independent Large Vocabulary Recognition Performance", Eurospeech, Madrid, pp. 1461-1464, 1995.
16. R. Gary Leonard, "A Database for Speaker-Independent Digit Recognition," ICASSP, 42.11.1-41.11.4, 1984.
17. J. C. R. Licklider and Irwin Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," JASA, 20(1), 42-51, 1948.
18. R. P. Lippmann, "Accurate Consonant Perception Without Mid-Frequency Speech Energy", IEEE Transactions on Speech and Audio Processing, 4(1), 66-69, 1996.
19. R. P. Lippmann, L. D. Braida, and N. I. Durlach, "A Study of Multichannel Amplitude Compression and Linear Amplification for Persons with Sensorineural Hearing Loss," JASA, 69, 524-534, 1981.
20. Liu, F.-H., et al., "Speech Recognition on Madarin Call Home: A Large-Vocabulary, Conversational, and Telephone Speech Corpus," ICASSP, 157-160, 1996.
21. M. Mack, J. Tierney, and M. E. T. Boyle, "The Intelligibility of Natural and LPC-Vocoded Words and Sentences Presented to Native and Non-Native Speakers of English," MIT Lincoln Laboratory, Tech. Report 869, 5 July 1990, DTIC AD-A226-180.
22. Alvin Martin, Personal Communication, 1996.
23. G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," JASA, 27, 338-352, 1955.
24. Miller, G., G. Heise, and W. Lichten, "The Intelligibility of Speech as a Function of the Context of the Test Materials," Journal of Experimental Psychology, 41(5), 329-335, 1951.
25. Miller, G.A. and J.C.R. Licklider, "The Intelligibility of Interrupted Speech." JASA, 22(2), 167-173, 1950.
26. Herve Bourlard, Hynek Hermansky, and Nelson Morgan, "Copernicus and the ASR Challenge - Waiting for Kepler," ARPA Speech Recognition Workshop, Harriman, NY, 1996.
27. Mullennix, J.W., D.B. Pisoni, and C.S. Martin, "Some effects of talker variability on spoken word recognition," JASA, 85, 365-378, 1989.
28. David S. Pallett, Jonathan G. Fiscus, et al., "1994 Benchmark Tests for the ARPA Spoken Language Program," SLST Workshop, Austin TX, 5-36, Morgan Kaufmann, 1995.
29. D. S. Pallett, "DARPA Resource Management and ATIS Benchmark Test Poster Session", Speech and Natural Language Workshop, Pacific Grove CA, 49-58, Morgan Kaufmann, 1991.
30. D. Paul, "The Design for the Wall Street Journal-Based CSR Corpus," DARPA Speech and Natural Language Workshop, 357-360, Morgan Kaufmann, 1992.
31. P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," ICASSP, 651-654, 1988.
32. Sommers, M. S.,Humes, L., & Pisoni, D. B., "The Effects of Speaking Rate and Stimulus Variability on Spoken Word Recognition by Young and Elderly Listeners," (Research on Spoken Language Processing, Progress Report No. 19). Speech Research Laboratory, Indiana University, 1993.
33. Woodland, P.C., et al., "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task," SLST Workshop, Harriman NY, Morgan Kaufmann, 1996.
34. Carl E. Williams and Michael H. L. Hecker, "Relation between Intelligibility Scores for Four Test Methods and Three Types of Speech Distortion," JASA, 44(4), 1002-1006, 1968.
35. S. J. Young, P. C. Woodland, and W. J. Byrne, "Spontaneous Speech Recognition for the Credit Card Corpus using the HTK Toolkit," IEEE Transactions on Speech and Audio Processing, 2(4), 615-621, 1994.
36. Zue, V., et al., "The MIT SUMMIT Speech Recognition System: A Progress Report," Speech and Natural Language Workshop Philadephia PA, 179-189, Morgan Kaufmann, 1989.